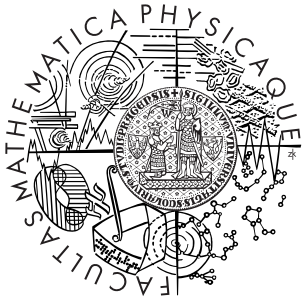
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁRSKA PRÁCA



Monika Jakubcová

Štatistická analýza cenzorovaných dát

Katedra pravdepodobnosti a matematickej štatistiky

Vedúci bakalárskej práce: Mgr. Petr Klášterecký

Študijný program: Matematika, Obecná matematika

2006

Rada by som sa na tomto mieste poďakovala vedúcemu bakalárskej práce Mgr. Petrovi Klášterecképu za jeho cenné rady, podnetné pripomienky a čas, ktorý mi venoval.

Prehlasujem, že som svoju bakalársku prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičaním práce a jej zverejňovaním.

V Prahe dňa 20. mája 2006 Monika Jakubcová

Obsah

Názov práce: Štatistická analýza cenzorovaných dát
Autor: Monika Jakubcová
Katedra : Katedra pravdepodobnosti a matematickej štatistiky
Vedúci bakalárskej práce: Mgr. Petr Klášterecký
e-mail vedúceho: klaster@karlin.mff.cuni.cz

Abstrakt: Predložená práca podáva prehľad o niektorých základných metódach štatistického spracovania cenzorovaných dát. V úvodnej kapitole uvedieme akým spôsobom cenzorované pozorovanie vzniká a popíšeme základné typy cenzorovania. Ďalej sa oboznámime s funkciami, ktoré cenzorované dáta popisujú a ich vzájomnými vzťahmi. Následne uvedieme neparametrické metódy odhadu týchto funkcií. Uvedené odhady neskôr použijeme pri konštrukcii intervalov spoľahlivosti a pri testovaní hypotéz. V záverečnej kapitole pripájame príklad, na ktorom demonštrujeme niektoré z popísaných metód a spôsob práce s cenzorovanými dátami v programe R.
Kľúčové slová: cenzorovanie, cenzoronané dáta, funkcia prežitia, Kaplanov-Meierov odhad, log-rank test

Title: Statistical analysis of censored data
Author: Monika Jakubcová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Petr Klášterecký
Supervisor's e-mail address: klaster@karlin.mff.cuni.cz

Abstract: The presented work gives an overview of some basic methods of statistical processing of censored data. In the opening article, we present the way a censored observation is obtained and we describe the basic types of censoring. Further, we get to know the basic quantities of censored data and their mutual relations. Afterwards, we introduce nonparametric estimations of basic quantities for censored data. We use the shown estimates later for the construction of confidence intervals and for hypotheses testing. In the final chapter, we add an example which demonstrates some of the aforementioned methods and the way censored data are handled in the R software.
Keywords: censoring, censored data, survival function, Kaplan-Meier estimator, log-rank test

Kapitola 1

Cenzorované dáta

1.1 Úvod

Cenzorované dáta sa vyskytujú v rôznych oblastiach, ako je napríklad medicína, biológia, epidemiológia, strojárstvo, ekonómia, atď. Môžme zhruba povedať, že cenzorované pozorovanie vzniká, ak je časť života jedinca presne známa a o zvyšku udalostí v jeho živote vieme povedať len toľko, že nastali v určitom intervale.

Poznáme niekoľko typov cenzorovania. Práca sa zameria na dáta cenzorované sprava (pojem bude podrobne vysvetlený nižšie), niektoré postupy v iných typoch cenzorovania spomenie len okrajovo. Bude sa venovať funkciám, ktoré cenzorované dáta popisujú a ich odhadom. Ďalej sa bude zaoberať intervalmi spoľahlivosti okolo a pozdĺž krivky prežitia. Kapitola 4 je venovaná testovaniu hypotéz či dáta majú niektoré zo známych rozdelení a porovnávaním viacerých súborov dát. Kapitola 5 obsahuje ilustratívny príklad.

1.2 Cenzorovanie

Cenzorované dáta sa vyskytujú veľmi často v medicíne, preto sa aj práca zameria práve na túto oblasť. Budeme hovoriť o vzorke n jedincov, u krorých je pozorovaná nejaká udalosť. Udalosťou môže byť napr. objavenie sa príznakov nejakej choroby, objavenie sa nádora, rozvinutie choroby, recidíva choroby, smrť atď. Alebo to môžu byť pozitívne udalosti, napr. zmiernenie príznakov po liečbe, narodenie dieťaťa atď.

Súbory cenzorovaných dát často vznikajú vtedy, keď je z časového alebo ekonomického hľadiska potrebné pokus zastaviť skôr ako je u všetkých jedincov pozorovaná udalosť. Môžu byť aj iné dôvody, pre ktoré niektoré údaje chýbajú a poznáme presný priebeh udalostí u jedinca len v určitom intervale. Pri spracovaní dát je žiadúce využiť aj informáciu, ktorú nesú cenzorované pozorovania.

Podľa spôsobu, akým cenzorované dáta vznikli, rozlišujeme cenzorovanie sprava, cenzorovanie zľava alebo intervalové cenzorovanie. Cenzorovanie sprava vzniká, ak vieme len to, že jedinec nezažil pozorovanú udalosť pred určitým časom, ale nevieme, ako sa jeho stav vyvíjal ďalej. Cenzorovanie zľava vzniká, ak vieme, že jedinec prežil udalosť už pred začiatkom štúdie, ale nemáme presný časový údaj, kedy udalosť nastala. Takéto údaje sa často získavajú napríklad z dotazníkov. A konečne intervalovo cenzorované dáta nás informujú, že udalosť nastala v časovom intervale. Intervalovo cenzorované dáta vznikajú napríklad pri pravidelnej kontrole pacienta. Príkladom môže byť ak zubár zistí zubný kaz, ale pri predchádzajúcej návšteve bol tento zub v poriadku.

1.2.1 Cenzorovanie sprava

Cenzorovanie sprava vzniká, ak nie je možné pozorovať všetkých jedincov až do výskytu udalosti. Väčšina práce sa bude zaoberať práve týmto cenzorovaním. Označme C_r dobu trvania šdúdie. Dáta v experimente, v ktorom vznikajú pozorovania cenzorované sprava, sú spravidla reprezentované pomocou dvojíc (T_i, δ_i), kde T_i je náhodná veličina a $\delta_i = 1$ značí, že udalosť nastala pred časom C_r, $\delta_i = 0$ značí, že nastalo cenzorovanie. Indikátor δ_i nazývame cenzorovací indikátor. Podľa mechanizmu cenzorovania rozlišujeme dva základné typy cenzorovania sprava.

Typ cenzorovania I

V tomto prípade dopredu zvolíme čas C_r. Udalosť je pozorovaná len v prípade, že nastala pred časom C_r. Pozorovanie prebieha na vzorke n jedincov. V praxi si tento postup často vynúti situácia, keď z časových dôvodov nezažijú udalosť všetky jedince alebo ak je pokus finančne náročný a je neekonomické nechať pokus plynúť, kým získame presné údaje o všetkých jedincoch. Všetky cenzorované pozorovania majú čas rovný dĺžke študovanej periódy, tj. C_r. Pri cenzorovaní sprava je n jedincov v štúdii a $X_1, \ldots, X_n$ sú časy do udalosti u každého jedinca. $X_1, \ldots, X_n$ sú nezávislé rovnako rozdelené náhodné veličiny s hustotou $f(x)$ a funkciou prežitia $S(x)$, (budeme

formálne definovať v kapitole 2). Presný čas X_i pre i-teho jedinca je známy len v prípade, že udalosť nastala pred časom C_r. Pre X_i väčšie ako C_r je i-te pozorovanie cenzorované v čase C_r.

Definícia: Náhodnú veličinu $T_i = min(X_i, C_r)$ nazývame *cenzorovaný čas.*

Ďalšími typmi sú tzv. progresívny typ cenzorovania I a zovšeobecnený typ cenzorovania I. Tieto typy sú podrobne popísané v [2].

Typ cenzorovania II

V štúdii je n jedincov. Štúdia sa zastaví, akonáhle nastane udalosť u prvých r jedincov, kde $r < n$ je dopredu určené prirodzené číslo. Typ cenzorovania II sa často používa pri testovaní vybavenia. Všetky časti vybavenia sú vložené do testu v rovnakom čase a test ukončíme, keď u r častí nastane udalosť. Takýto experiment môže ušetriť čas aj peniaze, pretože môže trvať veľmi dlho, kým dôjde k udalosti u všetkých častí vybavenia. Číslo r určuje počet častí, u ktorých došlo k udalosti a $n - r$ je počet cenzorovaných pozorovaní.

Ďalšími typmi sú tzv. progresívny typ cenzorovania II a zovšeobecnený typ cenzorovania II. Tieto typy sú podrobne popísané v [2].

1.2.2 Cenzorovanie zľava a intervalové cenzorovanie

Pre úplnosť zmienime tiež cenzorovanie zľava a intervalové cenzorovanie. Zbytok práce sa však bude zaoberať len cenzorovaním sprava.

Cenzorovanie zľava vzniká, ak vieme, že jedinec zažil pozorovanú udalosť niekedy pred časom C_1, ale nemáme presný údaj, kedy udalosť nastala. Presný čas do udalosti X_i je známy práve vtedy, keď $X_i \geq C_1$. Dáta cenzorované zľava sú reprezentované dvojicami náhodných veličin (T_i, ε_i). Ako pri cenzorovaní sprava, aj tu je $T_i = X_i$ ak bola udalosť pozorovaná v známom čase a $\varepsilon_i = 1$ značí, že udalosť nastala po čase C_1, $\varepsilon_i = 0$ značí, že nastalo cenzorovanie zľava. Pri cenzorovaní zľava náhodnú veličinu $T_i = max(X_i, C_1)$ nazývame cenzorovaný čas.

Ak je známy len údaj, že udalosť nastala v určitom intervale (L_i, R_i), ide o intervalové cenzorovanie. Táto situácia nastáva napríklad pri pravidelných kontrolách pacienta.

Kapitola 2

Charakteristiky cenzorovaných dát

V tejto kapitole sa oboznámime zo základnými charakteristikami cenzorovaných dát. Zatiaľ čo bežné štatistické metódy pracujú predovšetkým s hustotami a distribučnými funkciami náhodných veličin, tu sa častejšie používajú funkcie prežitia a rizikové funkcie. Čas do pozorovanej udalosti X je nezáporná náhodná veličina a jej rozdelenie popisujú predovšetkým štyri základné funkcie. *Funkcia prežitia*, ktorá vyjadruje pravdepodobnosť, že pozorovaná udalosť sa do času x nevyskytne; *riziková funkcia* vyjadrujúca riziko, že jedinec, ktorý do času x nezažil udalosť, zažije túto udalosť v nasledujúcom okamihu. Ďalej je to *hustota* náhodnej veličiny X a *stredná doba zostávajúceho života*, ktorá pre jedinca v čase x určuje očakávanú dĺžku zostávajúceho života. Ak poznáme jednu z týchto štyroch funkcií, zvyšné tri vieme jednoznačne odvodiť.

2.1 Funkcia prežitia

Definícia: Nech X je nezáporná náhodná veličina s distribučnou funkciou $F_X(x)$. Definujeme *funkciu prežitia* ako $S(x) = P(X > x)$. Funkcia prežitia je doplnok k distribučnej funkcii, teda $S(x) = 1 - F_X(x)$, kde

$$F_X(x) = P(X \leq x).$$

Ak je X spojitá náhodná veličina, potom $S(x)$ je spojitá funkcia. Funkcia prežitia je integrál z hustoty náhodnej veličiny $f_X(x)$, tj.:

$$S(x) = P(X > x) = \int_x^\infty f_X(t)dt.$$

Teda

$$f_X(x) = -\frac{dS(x)}{dx}.$$

Ak je X diskrétna náhodná veličina a nadobúda hodnoty $x_j, j = 1, 2, \ldots$ s pravdepodobnosťou $p(x_j) = P(X = x_j), \quad j = 1, 2, \ldots$, kde $x_1 < x_2 < \ldots$, potom je funkcia prežitia definovaná

$$S(x) = P(X > x) = \sum_{x_j > x} p(x_j)$$

a je po častiach konštantná funkcia. Je teda schodovitá, so skokmi veľkosti $p(x_j)$ v bodoch x_j.

Tvrdenie: Funkcia prežitia $S(x)$ náhodnej veličiny X má nasledujúce vlastnosti:
(i) $S(x)$ je nerastúca funkcia
(ii) $\lim_{t\to\infty} S(t) = 0, \qquad S(0) = 1.$

Dôkaz. Tvrdenie vyplýva z vlastností distribučnej funkcie nezápornej náhodnej veličiny.

2.2 Riziková funkcia

Definícia: Nech X je náhodná veličina. Funkcia

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x | X \ge x)}{\Delta x}$$

sa nazýva *riziková funkcia.*

Riziková funkcia je vždy nezáporná $h(x) \ge 0$. Výraz $h(x)\Delta x$ môžme interpretovať ako približnú pravdepodobnosť, že udalosť nastane v okamihu nasledujúcom po čase x za podmienky, že udalosť u jedinca ešte nenastala.

Riziková funkcia môže mať rôzny priebeh. Príkladom modelu s rastúcou rizikovou funkciou je, ak predpokladáme starnutie pacientov alebo opotrebovanie orgánov. S klesajúcou rizikovou funkciou sa môžme stretnúť ak sa dáta týkajú napríklad pacientov po translpantácii orgánu.

Ak je X spojitá náhodná veličina potom

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d\ln[S(x)]}{dx}.$$

Ak je X diskrétna náhodná veličina, potom je riziková funkcia definovaná nasledovne:

$$h(x_j) = P(X = x_j | X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})}, \quad j = 1, 2, \ldots$$

kde $S(x_0) = 1$ a $p(x_j) = S(x_{j-1}) - S(x_j)$. Z toho vyplýva, že

$$h(x_j) = 1 - \frac{S(x_j)}{S(x_{j-1})}, \quad j = 1, 2, \ldots$$

S rizikovou funkciou úzko súvisí kumulatívne riziko, ktoré značíme $H(x)$ a definujeme

$$H(x) = \int_0^x h(u)du = -\ln[S(x)]. \tag{2.1}$$

Úpravou získame pre funkciu prežitia vzťah

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)du\right].$$

Funkcia prežitia môže byť vyjadrená aj ako súčin podmienených pravdepodobností prežitia

$$S(x) = \prod_{x_j \leq x} \frac{P(X = x_j | X \geq x_j) P(X = x_{j+1})}{P(X = x_{j+1} | X \geq x_{j+1}) P(X = x_j)} = \prod_{x_j \leq x} \frac{S(x_j)}{S(x_{j-1})}.$$

Ak použitím výraz $\frac{S(x_j)}{S(x_{j-1})} = 1 - h(x_j)$, po dosadení dostaneme funkciu prežitia vyjadrenú pomocou rizikovej funkcie pre diskrétnu náhodnú veličinu X

$$S(x) = \prod_{x_j \leq x} [1 - h(x_j)].$$

Definícia: Nech X je náhodná veličina. Funkcia $mrl(x) = E(X - x | X > x)$ sa nazýva *stredná doba zostávajúceho života*. Číslu $\mu = mrl(0)$ sa hovorí *stredná doba života*.

Pre spojitú náhodnú veličinu platí

$$mrl(x) = \frac{\int_x^\infty (t-x)f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)}$$

a

$$\mu = E(X) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt.$$

Táto funkcia určuje pre jedinca v čase x očakávanú dĺžku zostávajúceho života. Je definovaná ako podiel plochy pod krivkou prežitia od času x a $S(x)$. Stredná doba života $\mu = mrl(0)$ je zrejme plocha pod celou krivkou prežitia.

Kapitola 3

Odhady

V tejto kapitole sa budeme venovať najpoužívanejším odhadom funkcie prežitia a kumulatívneho rizika. Tieto odhady sa použijú napríklad pri testovaní hypotéz uvedených v kapitole 4. Ďalej sa kapitola zameria na intervaly spoľahlivosti okolo a pozdĺž kriviky prežitia.

3.1 Odhady funkcie prežitia a kumulatívneho rizika pre dáta cenzorované sprava

Predpokladáme, že udalosť sa vyskytuje v čase X, kde X je náhodná veličina. Zameriame sa na vzorku sprava cenzorovaných dát. Dáta sa skladajú z n dvojíc údajov (T_i, δ_i), definovaných v kapitole 2. Budeme predpokladať, že čas cenzorovania a čas, kedy udalosť nastala, sú nezávislé. Popísané metódy sú vhodné pre všetky bežné typy cenzorovania uvedené v kapitole 1.

Značenie: Časy, v ktorých sa vyskytla aspoň jedna udalosť, značíme $t_1, \ldots, t_D$, $D \in N$, kde $t_1 < t_2 < \ldots < t_D$. Označme d_i počet udalostí, ktoré nastali v čase t_i. Ďalej označme Y_i počet jedincov, ktorí sú v čase t_i v riziku, tj. udalosť ani cenzorovanie u nich ešte nenastalo.

Hodnota $\frac{d_i}{Y_i}$ dáva odhad podmienenej pravdepodobnosti, že jedinec zažije udalosť v čase t_i, za podmienky, že udalosť ani cenzorovanie u neho pred časom t_i nenastalo. Táto hodnota je základom pri odvodzovaní odhadu funkcie prežitia a kumulatívneho rizika.

Definícia: *Kaplanov-Meierov odhad funkcie prežitia* je definovaný vzťahom

$$\hat{S}(t) = \begin{cases} 1 & \text{ak} \quad t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{ak} \quad t_1 \leq t, \end{cases}$$

kde t_1 je čas, kedy nastala prvá udalosť.

Tento odhad funkcie prežitia navrhli Kaplan a Meier v roku 1958. Cenzorovanie sa tu prejaví v hodnote Y_i. Ak by sme cenzorované dáta nezapočítali, potom $Y_i = n - \sum_{j=1}^{i} d_j$ teda, ak sú medzi dátami cenzorované pozorovania potom $Y_i \leq n - \sum_{j=1}^{i} d_j$. Vo výbere bez cenzorovaných pozorovaní sa Kaplanov-Meierov odhad redukuje na doplnok empirickej distribučnej funkcie.

Kaplanov-Meierov odhad je po častiach konštantná funkcia. Skoky nastávajú v časoch, kedy nastala udalosť. Veľkosť skoku závisí na počte pozorovaných udalostí v čase t_i a na počte cenzorovaných pozorovaní pred časom t_i. Cenzorované pozorovania sa prejavujú na odhade krivky prežitia najmä v jej neskorších častiach, a to tak, že sa zväčšujú veľkosti skokov. Kaplanov-Meierov odhad je konzistentným odhadom $S(t)$ a jeho rozptyl vyjadruje Greenwoodov vzorec:

$$\hat{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}. \tag{3.1}$$

Smerodajná odchýlka Kaplanovho-Meierovho odhadu je $[\hat{V}[\hat{S}(t)]]^{\frac{1}{2}}$. Kaplanov-Meierov odhad má za pomerne obecných predpokladov asymptoticky normálne rozdelenie, čo využijeme neskôr pri konštrukcii intervalov spoľahlivosti.

Kaplanov-Meierov odhad dáva odhad funkcie prežitia. Zo vzájomného vzťahu kumulatívneho rizika a funkcie prežitia (2.1) dostávame nasledujúci odhad kumulatívneho rizika

$$\hat{H}(t) = -\ln[\hat{S}(t)]$$

Iný odhad kumulatívneho rizika je definovaný predpisom

$$\tilde{H}(t) = \begin{cases} 0 & \text{ak} \quad t \leq t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{ak} \quad t_1 \leq t \end{cases} \tag{3.2}$$

a nazýva sa *Nelsonov-Aalenov odhad kumulatívneho rizika.* Podľa [2] je jeho použitie vhodné najmä pri výberoch malého rozsahu. Asymptoticky sú odhady $\hat{H}(t)$ a $\tilde{H}(t)$ ekvivalentné.

Rozptyl tohto odhadu je

$$\sigma_H^2(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}$$

Aj v tomto prípade môžme použiť vzťah medzi kumulatívnym rizikom a funkciou prežitia (2.1) a dostaneme inú variantu odhadu funkcie prežitia $\tilde{S}(t) = \exp[-\tilde{H}(t)]$.

3.2 Intervaly spoľahlivosti okolo krivky prežitia

Interval spoľahlivosti okolo krivky v bode $S(t_0)$ je interval spoľahlivosti v pevne zvolenom čase t_0. Intervaly spoľahlivosti sú založené na Kaplanovom-Meierovom odhade a jeho smerodajnej odchýlke z predošlej časti. Pôjde o obojstranný interval spoľahlivosti pre $S(t_0)$ na hladine $1 - \alpha$. Teda pravdepodobnosť, že interval prekryje skutočnú hodnotu funkcie prežitia v čase t_0 je rovná $1 - \alpha$. Číslo α sa spravidla volí malé. Najčastejšie $\alpha = 0,01$ alebo $\alpha = 0,05$. Spoľahlivosť sa často udáva v percentách (95% alebo 99%).

Pre stručnejší zápis označíme sumu z Greenwoodovho vzorca (3.1)

$$\sigma_S^2(t) = \frac{\hat{V}[\hat{S}(t)]}{[\hat{S}(t)]^2}.$$

Najčastejšie sa používa lineárny interval spoľahlivosti pre $S(t_0)$ na hladine $1 - \alpha$.

$$(\hat{S}(t_0) - Z_{1-\frac{\alpha}{2}}\sigma_S(t_0)\hat{S}(t_0), \hat{S}(t_0) + Z_{1-\frac{\alpha}{2}}\sigma_S(t_0)\hat{S}(t_0)),$$

kde $Z_{1-\frac{\alpha}{2}}$ je $1-\frac{\alpha}{2}$ kvantil štandardizovaného normálneho rozdelenia $N(0,1)$. Tieto kvantily je môžme nájsť v bežných štatistických tabuľkách.

Práve lineárny tvar intervalu spoľahlivosti ponúka aj väčšina softwarových balíkov. V príklade v kapitole 5 ukážeme výpočet tohoto intervalu v programe R.

Iné intervaly spoľahlivosti vznikajú ak Kaplanov-Meierov odhad $\hat{S}(t_0)$ transformujeme. Najčastejšie sa používajú logaritmické transformácie a arcsin-transformácie. Príslušné tvary transformovaných intervalov sa dajú nájsť v [2].

Logaritmicky transformovaný interval aj arcsin-transformovaný interval nie sú, na rozdiel od lineárneho intervalu spoľahlivosti, symetrické okolo

bodového odhadu funkcie prežitia. Podľa [2] majú tieto intervaly v niektorých prípadoch, napríklad pri malom počte pozorovaní, lepšie vlastnosti. Lineárny interval spoľahlivosti je založený na asymptotickej normalite odhadu $\hat{S}(t)$. Táto aproximácia je pri malých výberoch nepresná a rozdelenia bývajú zošikmené. To sa dá čiastočne odstrániť práve transformáciou.

3.3 Intervaly spoľahlivosti pozdĺž krivky prežitia

Intervaly spoľahlivosti v predošlej časti sú platné len pre pevne zvolený čas t_0. V tejto časti popíšeme intervaly spoľahlivosti pozdĺž krivky prežitia $S(t)$ pre všetky časy z intervalu (t_L, t_U). Teda nájdeme hornú a dolnú hranicu pásu spoľahlivosti, ktorý prekryje funkciu prežitia pre všetky časy t z intervalu (t_L, t_U) s pravdepodobnosťou $1 - \alpha$.

Definícia: Horná a dolná hranica pásu spoľahlivosti sú náhodné funkcie $L(t)$ a $U(t)$, pre ktoré platí

$$P[L(t) \leq S(t) \leq U(t), \text{ pre všetky časy t také, že } t_L \leq t \leq t_U] = 1 - \alpha.$$

Interval $[L(t), U(t)]$ nazývame *interval spoľahlivosti pozdĺž krivky prežitia* $S(t)$ na hladine $1 - \alpha$.

Môžme zvoliť dva spôsoby ako pristupovať k týmto intervalom. Prvý spôsob navrhol Nail v roku 1984. Tu treba vytvoriť interval spoľahlivosti pozdĺž krivky primeraný k intervalovým odhadom okolo krivky z predošlej časti. Tieto intervaly sa nazývajú EP pásy spoľahlivosti. Položme $t_L < t_U$, kde t_L je väčší alebo rovný najmenšiemu cenzorovanému času a t_U je menší alebo rovný najväčšiemu cenzorovanému času. Pre vzorku n dát a funkciu prežitia $S(t)$ definujeme

$$a_L = \frac{n\sigma^2{}_S(t_L)}{1 + n\sigma^2{}_S(t_L)}$$

a

$$a_U = \frac{n\sigma^2{}_S(t_U)}{1 + n\sigma^2{}_S(t_U)}.$$

Požadujeme $0 < a_L < a_U < 1$. Pri zostavovaní intervalu spoľahlivosti pozdĺž $S(t)$ na hladine $1 - \alpha$ pre časový interval (t_L, t_U) musíme nájsť koeficienty $c_\alpha(a_L, a_U)$, ktoré sú uvedené v tabuľkách, napr. v literatúre [2]. Podobne ako pri intervaloch spoľahlivosti okolo $S(t)$ aj tu môžme uviesť tri typy obojstranných intervalov spoľahlivosti na hladine $1 - \alpha$ pozdĺž $S(t)$ a pre časový interval (t_L, t_U).
Lineárny interval spoľahlivosti má tvar

$$(\hat{S}(t) - c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t), \hat{S}(t) + c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t)),$$

ďalšie dva tvary intervalov spoľahlivosti vznikajú logaritmickou transformáciou a arcsin-transformáciou. Príslušné tvary transformovaných intervalov sa opäť dajú nájsť v [2].

Druhý spôsob ako môžme pristupovať k týmto intervalom navrhli Hall a Wellner v roku 1980. Tieto intervaly spoľahlivosti nie sú primerané k intervalovým odhadom okolo krivky z predošlej časti. Tu pripúšťame hodnotu $t_L = 0$. V tomto prípade potrebujeme na zostavenie intervalov spoľahlivosti pozdĺž $S(t)$ na hladine $1-\alpha$ pre časový interval (t_L, t_U) koeficienty $k_\alpha(a_L, a_U)$, ktoré sú taktiež uvedené v tabuľkách, napr. v [2]. Znovu môžme uviesť tri typy týchto intervalov.
Lineárny interval spoľahlivosti má tvar

$$\left(\hat{S}(t) - \frac{k_\alpha(a_L, a_U)[1 + n\sigma^2{}_S(t)]}{n^{\frac{1}{2}}}\hat{S}(t), \hat{S}(t) + \frac{k_\alpha(a_L, a_U)[1 + n\sigma^2{}_S(t)]}{n^{\frac{1}{2}}}\hat{S}(t)\right),$$

tvary intervalov spoľahlivosti vzniknutých logaritmickou transformáciou a arcsin-transformáciou sa dajú nájsť v [2].

Kapitola 4

Testovanie hypotéz

V tejto kapitole budeme uvažovať jedno-, dvoj- a viacvýberové testy. Pri konštrukcii testových štatistík využijeme najmä Nelsonov-Aalenov odhad kumulatívneho rizika (3.2) z kapitoly 3.

4.1 Jednovýberové testy

Pracujeme s jednou vzorkou, krorá obsahuje n dvojíc (t_i, δ_i). Z týchto údajov urobíme Nelsonov-Aalenov odhad kumulatívneho rizika (3.2). Prejdeme k rizikovej funkcii použitím vzťahu medzi kumulatívnym rizikom a rizikovou funkciou $h(x) = \frac{dH(x)}{dx}$. Testy sú založené na porovnávaní rizikovej funkcie odhadnutej z dát $h(t)$ a teoretickej rizikovej funkcie $h_0(t)$. Nulová hypotéza hovorí, že dáta pochádzajú z rozdelenia so známou rizikovou funkciou $h_0(t)$. Pri posudzovaní je možné dávať viac váhy na určitú časť krivky.

Formálne zapísané, chceme testovať hypotézu o rizikovej funkcii
$H_0 : h(t) = h_0(t)$ pre všetky $t \leq \tau$, proti alternatíve
$A : h(t) \neq h_0(t)$ pre nejaké $t \leq \tau$.

Teoretická riziková funkcia $h_0(t)$ je presne definovaná v intervale $[0, \tau]$. Najčastejšie bude pre nás τ najväčší pozorovaný čas. Za platnosti nulovej hypotézy očakávame, že riziko v čase t_i je $h_0(t_i)$ pre všetky $t_i \in [0, \tau]$.

Tvrdenie: Nech W(t) je váhová funkcia taká, že platí W(t)=0 vždy, keď Y(t)=0.

Testová štatistika $\frac{(Z(\tau))^2}{V[Z(\tau)]}$, kde

$$Z(\tau) = \sum_{i=1}^{D} W(t_i)\frac{d_i}{Y(t_i)} - \int_0^\tau W(s)h_0(s)ds$$

a

$$V[Z(\tau)] = \int_0^\tau [W(s)]^2 \frac{h_0(s)}{Y(s)}ds$$

má za platnosti nulovej hypotézy asymptoticky rozdelenie χ_1^2.

Tvrdenie sme bez dôkazu prevzali z [2].

Kritický obor testu je

$$\frac{(Z(\tau))^2}{V[Z(\tau)]} \geq \chi_1^2(1-\alpha),$$

kde $\chi_1^2(1-\alpha)$ je $(1-\alpha)$ kvantil χ^2 rozdelenia o 1 stupni voľnosti. Hypotézu H_0 zamietame pri veľkých hodnotách testovej štatistiky.

Poznámka: Na testovenie hypotézy $H_0 : h(t) = h_0(t)$ proti jednostrannej alternatíve, $A : h(t) > h_0(t)$ je možné použiť štatistiku $\frac{Z(\tau)}{V[Z(\tau)]^{\frac{1}{2}}}$. Táto štatistika má za platnosti nulovej hypotézy asymptoticky štandardizované normálne rozdelenie $N(0,1)$. Nulová hypotéza je znovu zamietnutá pri veľkej hodnote štatistiky.

Voľba váhovej funkcie

Najčastejšie sa používa váhová funkcia $W(t) = Y(t)$. Pri tejto voľbe dostávame tzv. *jednovýberový log-rank test.* Iná možnosť voľby váhovej funkcie je zvoliť niektorú zo skupiny Harringtonových-Flemingových váhových funkcií $W_{HF}(t) = Y(t)S_0(t)^p[1 - S_0(t)]^q, p \geq 0, q \geq 0$, kde $S_0(t) = \exp[-H_0(t)]$ je teoretická funkcia prežitia pri nulovej hypotéze. Voľbou p a q môžme ovplyvniť, v ktorej časti rizikovej funkcie sú odchýlky od nulovej hypotézy závažnejšie. Ak chceme dať viac váhy na začiatočné odchýlky od nulovej hypotézy, volíme p oveľa väčšie ako q, ak chceme dať viac váhy na neskoršie odchýlky od nulovej hypotézy, volíme p naopak oveľa menšie ako q. Voľbou $p = q > 0$ dávame najväčšiu váhu do strednej časti rizivovej funkcie odhadnutej z dát a teoretickej rizikovej funkcie. Špeciálnou voľbou $p = q = 0$ dostávame tzv. log-rank test.

4.2 Dvoj- a viacvýberové testy

V predošlej časti sme sa venovali jednovýberovým testom, ktoré robili vážené porovnanie rizikovej funkcie odhadnutej z dát a teoretickej rizikovej funkcie. Teraz budeme porovnávať dáta, ktoré pozostávajú z K, $K \geq 2$ nezávislých sprava cenzorovaných vzoriek. Pre vzorku K_i máme odhadnutú rizikovú funkciu $h_i(t)$. Opäť používame Nelsonov-Aalenov odhad kumulatívneho rizika (3.2) a vzťah medzi kumulatívnym rizikom a rizikovou funkciou. Budeme testovať hypotézu

$$H_0 : h_1(t) = h_2(t) = \ldots = h_K(t) \qquad \text{pre} \quad \forall t \leq \tau,$$

proti alternatíve

$$A : \text{aspoň jedna} \quad h_j(t) \quad \text{je odlišná od ostatných pre nejaké} \quad t \leq \tau.$$

V tomto prípade obyčajne za τ berieme najväčší čas, v ktorom majú všetky skupiny aspoň jedného jedinca v riziku. Nulovú hypotézu v tomto prípade zamietame, ak je aspoň jedna vzorka v rozpore s ostatnými v nejakom čase.

Značenie: Označme $t_1 < t_2 < \ldots < t_D$ časy udalostí v súhrne všetkých vzoriek. Číslo d_{ij} je počet udalostí v j-tej vzorke pozorovaných v čase t_i. Y_{ij} je počet jedincov v riziku v j-tej vzorke v čase t_i, $j = 1, \ldots, K$, $i = 1, \ldots, D$. Ďalej označme $d_i = \sum_{j=1}^{K} d_{ij}$ počet udalostí v čase t_i a $Y_i = \sum_{j=1}^{K} Y_{ij}$ je počet jedincov v riziku v súhrne všetkých K vzoriek v čase t_i, $i = 1, \ldots, D$.

Test H_0 je založený na porovnaní vážených rizikových funkcií odhadnutých z dát $h_j(t)$. Za platnosti nulovej hypotézy je riziková funkcia odhadnutá z dát v j-tej vzorke rovná rizikovej funkcii odhadnutej zo súhrnu všetkých vzoriek, teda $\frac{d_i}{Y_i}$. Ak použijeme dáta len z j-tej vzorky, odhad rizikovej funkcie je $\frac{d_{ij}}{Y_{ij}}$.

Nech $W_j(t)$ je kladná váhová funkcia taká, že platí $W_j(t_i) = 0$ vždy, keď $Y_{ij} = 0$. Test H_0 je založený na štatistikách

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i) \left\{ \frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right\}, \quad j = 1, \ldots, K.$$

Ak je štatistika $Z_j(\tau)$ blízko nuly pre každé $j = 1, \ldots, K$, potom je len malá šanca, že nulová hypotéza bude zamietnutá. Naopak, ak je jedna

zo štatistík $Z_j(\tau)$ ďaleko od nuly, potom je pravdepodobné, že táto vzorka má rizikovú funkciu v rozpore s tou, ktorú sme očakavali pri platnosti nulovej hypotézy.

Voľba váhovej funkcie

Napriek tomu, že sa väčšinou vyhýbame situácii, že váhové funkcie sú rôzne pre každé $Z_j(\tau)$, v praxi sa obyčajne používajú testy s váhovou funkciou $W_j(t_i) = Y_{ij}W(t_i)$. Váha $W(t_i)$ je pre všetky vzorky rovnaká a Y_{ij} je počet jedincov v riziku v j-tej vzorke v čase t_i. Teda pre tento výber váhovej funkcie dostaneme testovú štatistiku

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i) \left\{ d_{ij} - Y_{ij}\frac{d_i}{Y_i} \right\}, \quad j = 1, \ldots, K. \tag{4.1}$$

Pri použití tejto triedy váhových funkcií je testová štatistika suma z váženého rozdielu medzi pozorovaným počtom udalostí a očakávaným počtom udalostí v j-tej vzorke. Očakávaný počet udalostí v j-tej vzorke v čase t_i je pomer jedincov v riziku $\frac{Y_{ij}}{Y_i}$, ktorí sú vo vzorke j v čase t_i vynásobený počtom udalostí v čase t_i.

$Z_j(\tau)$ vo vzorci (4.1) má rozptyl

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} (W(t_i))^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i}\right) \left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad j = 1, \ldots, K$$

a kovariancia $cov(Z_j(\tau), Z_g(\tau))$ je rovná

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} (W(t_i))^2 \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad g \neq j.$$

Zložky vektora $(Z_1(\tau), \ldots, Z_K(\tau))$ sú lineárne závislé, pretože $\sum_{j=1}^{K} Z_j(\tau) = 0$. Kvôli závislosti $Z_j(\tau)$ skúmame len $K-1$ z nich.

Tvrdenie: Nech Σ je kovariančná matica typu $(K-1) \times (K-1)$ odhadnutá z dát s prvkami $\hat{\sigma}_{jg}$. Potom je testová štatistika T^2 daná kvadratickou formou

$$T^2 = (Z_1(\tau), \ldots, Z_{K-1}(\tau))\Sigma^{-1}(Z_1(\tau), \ldots, Z_{K-1}(\tau))^T$$

má za platnosti nulovej hypotézy asymptoticky rozdelenie χ^2_{K-1}.

Tvrdenie sme bez dôkazu prevzali z [2].

Kritický obor testu je teda

$$T^2 \geq \chi^2_{K-1}(\alpha)$$

hypotézu H_0 zamietame pri veľkých hodnotách testovej štatistiky.

Voľba váhovej funkcie

Vo väčšine prípadov volíme váhovú funkciu $W(t) = 1$ pre všetky t. Táto voľba váhovej funkcie vedie k tzv. log-rank testu. Gehan v roku 1965 navrhol váhovú funkciu $W(t_i) = Y_i$. Táto voľba vedie k tzv. zovšeobecnenému dvojvýberovému Mannovmu-Whitneyovmu-Wilcoxonovmu testu a k tzv. Kruskalovmu-Wallisovmu testu. Tarone a Ware v roku 1977 navrhli triedu testov s váhovými funkciami $W(t_i) = f(Y_i)$, kde funkcia f je pevne zvolená. Vybrali z tejto triedy váhových funkcií funkciu $f(y) = y^{\frac{1}{2}}$. Táto trieda váhových funkcií dáva väčšiu váhu na rozdiely medzi pozorovaným počtom udalostí a očakávaným počtom udalostí v j-tej vzorke v časoch, kde je najviac dát.

Kapitola 5

Príklad

V tejto kapitole uvedieme ilustratívny príklad, na ktorom demonštrujeme metódy popísané v predchádzajúcich kapitolách. Pri spracovaní príkladu sa pozrieme, aké možnosti nám poskytuje program R.

5.1 Dáta

V použitej štúdii sú pacienti trpiaci rakovinou močového mechúra. Po zákroku, pri ktorom im bol nádor odstránený, boli pacienti náhodne rozdelení do dvoch skupín. Jednej skupine bol podávaný liek s neúčinnou látkou a druhej skupine podávali liek s chemoterapeutickou látkou (thiotepa). Pozorovali sa časy do recidívy nádora u oboch skupín pacientov. Budeme skúmať vplyv jednotlivých liečebných postupov na čas do prvej recidívy nádora. Dáta k tejto štúdii sú uvedené v [1] a pozostávajú z nasledujúcich údajov:
Time: čas do udalosti v mesiacoch
Stav: stav pacienta (0=cenzorovené pozorovanie, 1=recidíva nádora)
Treat: typ liečby (1=neúčinná látka, 2=thiotepa).

5.2 Software

Program R je voľne šíriteľný štatistický software. Je možné ho získať na adrese www.r-project.org. V programe R použijeme tri základné funkcie na spracovanie cenzorovaných dát. Technická funkcia *Surv* vytvorí vstup do ďalších funkcií. Funkcia *survfit* odhaduje krivku prežitia pre cenzorované dáta a funkcia *survdiff* robí testy hypotéz. Pri spustení programu

sa tieto funkcie spravidla nenačítajú automaticky, preto je potrebné ich načítať z knižnice pomocou príkazu *library(survival)*.

Funkcia *Surv* spracuje dvojice (T_i, δ_i) pre ďalšie použitie
Možnosť použitia: Surv(time, delta), kde

time: je cenzorovaný čas,
delta: je cenzorovací indikátor.

Funkcia *survfit* počíta odhad krivky prežitia pre cenzorované dáta. Používa buď Kaplanovu-Meierovu alebo Flemingovu-Harringtonovu metódu.
Možnosť použitia: survfit(formula, subset, conf.int=.95, type=c("kaplan- meier"), error=c("greenwood"), conf.type=c("log","log-log","plain","none"))

V časti *formula* zadávame výstup z funkcie *Surv*, pomocou znaku $\sim$ môžme dáta rozdeliť podľa nejakého kritéria a funkcia *survfit* počíta krivku prežitia pre každú skupinu zvlášť.

Voľba *conf.type* ovplyvňuje typ intervalu spoľahlivosti okolo krivky prežitia. Voľba "plain" počíta lineárny tvar intervalu spoľahlivosti, voľba "log-log" počíta logaritmický tvar intervalu spoľahlivosti.

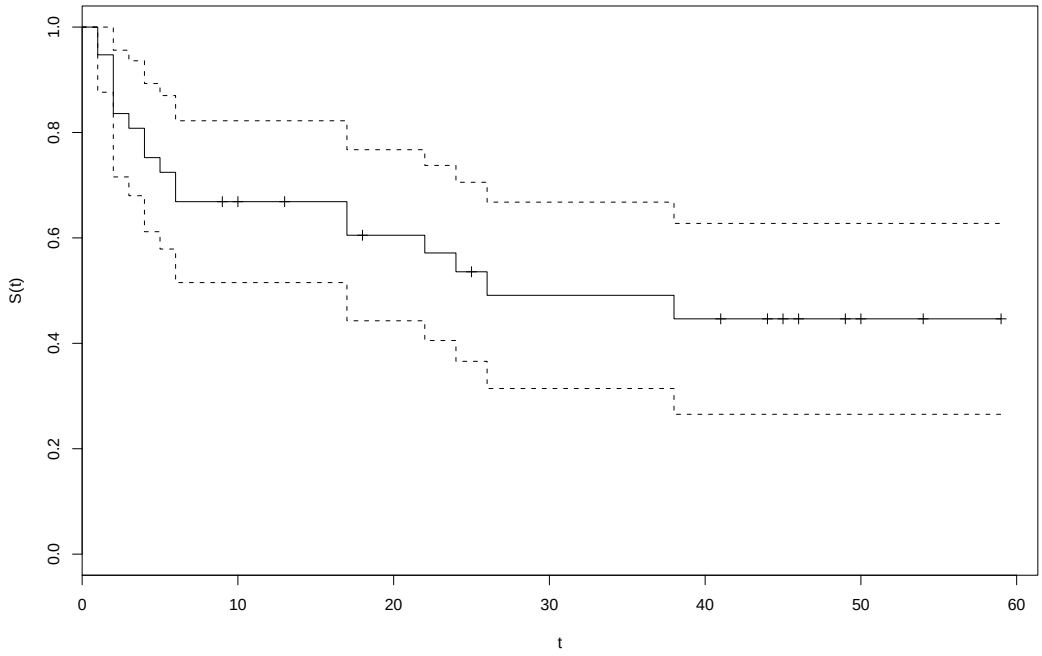
Funkcia *survdiff* testuje rozdiely medzi dvoma alebo viacerými krivkami prežitia, alebo robí test pre jednu krivku prežitia odhadnutú z dát proti teoretickej krivke.
Možnosť použitia: survdiff(formula, rho=0)

Parameter *rho* určuje typ testu. Voľba rho=0 určuje log-rank test.

5.3 Výsledky

Najprv budeme počítať lineárny interval spoľahlivosti okolo krivky prežitia pre skupinu pacientov, ktorým podávali liek s účinnou látkou thiotepa. Na obrázku 1 je zakreslený Kaplanov-Meierov odhad funkcie prežitia. Krivka je po častiach lineárna a skok nastáva v čase, kde nastala udalosť. Prerušovanou čiarou je znázornený obojstranný lineárny interval spoľahlivosti okolo krivky prežitia na hladine 95%. Za obrázkom je pripojená tabuľka 1, v ktorej sú uvedené hodnoty odhadov pre pacientov, ktorí užívali účinnú látku.

Obrázok 1: Kaplanov-Meierov odhad krivky prežitia pre typ liečby thiotepa



Tabuľka 1: Výstup funkcie summary(fit2) pre typ liečby thiotepa

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	38	2	0.947	0.0362	0.876	1.000
2	34	4	0.836	0.0613	0.716	0.956
3	30	1	0.808	0.0653	0.680	0.936
4	29	2	0.752	0.0717	0.612	0.893
5	27	1	0.724	0.0743	0.579	0.870
6	26	2	0.669	0.0783	0.515	0.822
17	21	2	0.605	0.0828	0.443	0.767
22	18	1	0.571	0.0848	0.405	0.738
24	16	1	0.536	0.0867	0.366	0.706
26	12	1	0.491	0.0902	0.314	0.668
38	11	1	0.446	0.0924	0.265	0.628

Tabuľka 1 v jednotlivých stĺpcoch udáva čas pozorovania t_i, počet jedincov v riziku Y_i, počet udalostí d_i, odhat $\hat{S}(t_i)$, odhad rozptylu a 95% interval spoľahlivosti v bode t_i.

Ďalej prevedieme log-rank test, aby sme zistili, či sa krivky líšia, ak dáta rozdelíme podľa spôsobu liečby pacientov. Výstup programu R obsahuje tabuľka 2, ktorá pre jednotlivé spôsoby liečby udáva najmä počet pozorovaní a pozorovanú a očakávanú početnosť.

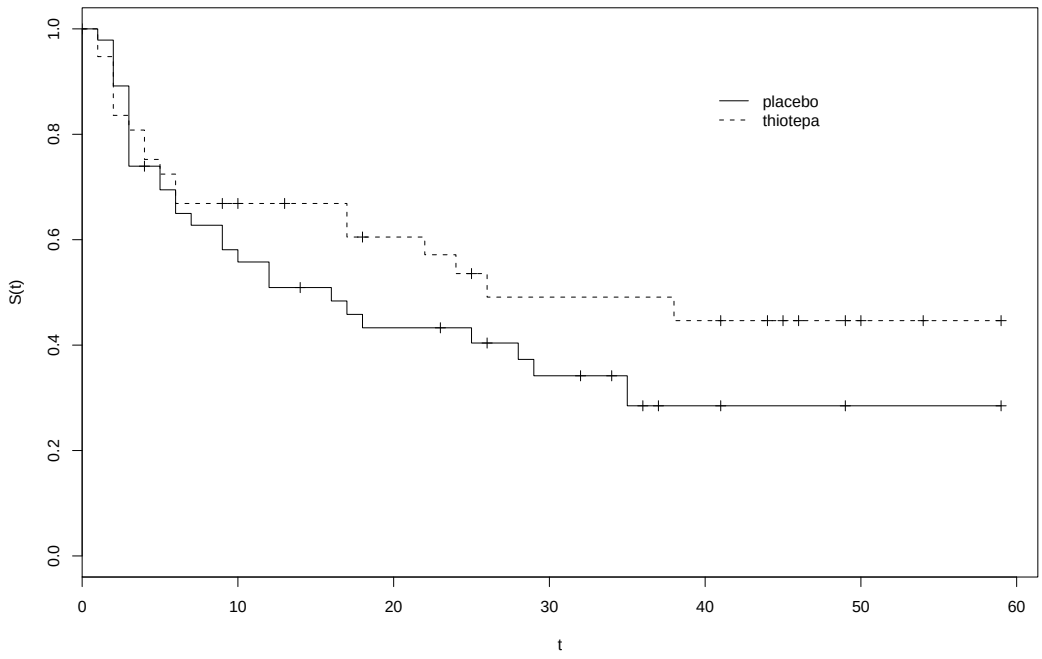
Tabuľka 2: Výstup funkcie print(test)

Treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Treat=1	48	29	24.9	0.671	1.52
Treat=2	38	18	22.1	0.757	1.52

Hodnota α kvantilu ${\chi^2}_1$ rozdelenia je 1,5. Výsledok testu $p = 0.217$ značí, že na bežnej hladine 5% nezamietame hypotézu o rovnosti kriviek. Účinok látky thiotepa výrazne neovplyvnil výskyt prvej recidívy nádora.

Na obrázku 2 je zakreslený Kaplanov-Meierov odhad funkcie prežitia pre obe skupiny pacientov. Je vidieť, že u skupiny, ktorá užívala účinnú látku, je odhad krivky prežitia väčší, ako u pacientov, ktorí užívali neúčinnú látku. Z obrázku by sme mohli usúdiť, že látka thiotepa má priaznivý vplyv na priebeh choroby. Test však tento priaznivý vplyv nepreukázal.

Obrázok 2: Porovnanie dvoch druhov liečby



Dadatok

Príkazy v R použité v kapitole 5.

```
>Stime<-Surv(Time, delta)
>fit1<-survfit(Stime, subset=Treat==1, type="kaplan-meier",
conf.type="plain")
>fit2<-survfit(Stime, subset=Treat==2, type="kaplan-meier",
conf.type="plain")
>summary(fit2)
>plot(fit2, ylab="S(t)", xlab="t")
>fit3<-survfit(Stime~Treat, type="kaplan-meier",
conf.type="plain")
>plot(fit3, lty=c(1,2), ylab="S(t)", xlab="t")
>legend(40,.9, legend=c("placebo","thiotepa"), lty=c(1,2),
bty="n")
>test<-survdiff(Stime~Treat, rho=0)
>print(test)
```

Literatura

[1] Collett, David: *Modelling Survival Data in Medical Research*, Chapman & Hall/CRC, 2003.

[2] Klein, J.P., Moeschberger, M. L.: *Survival Analysis*, Springer, 2003.